Three New Ways Used to Improve the Data

1. Feature Engineering: Created a new feature called 'total_guests' which is a sum of 'children' and 'babies', potentially capturing family bookings better.

2. Handling Missing Values: Filled missing values with the median, which is robust to outliers and can improve model accuracy.

3. Normalization: Applied MinMax scaling to numerical features to ensure that all features contribute equally to the result.

Error Analysis and Data Improvements

1. One of the key steps in my data-centric AI approach was to conduct a thorough error analysis. The confusion matrix, as shown below, was instrumental in this process. The confusion matrix reveals the number of true positives, true negatives, false positives, and false negatives. Specifically, my model predicted:

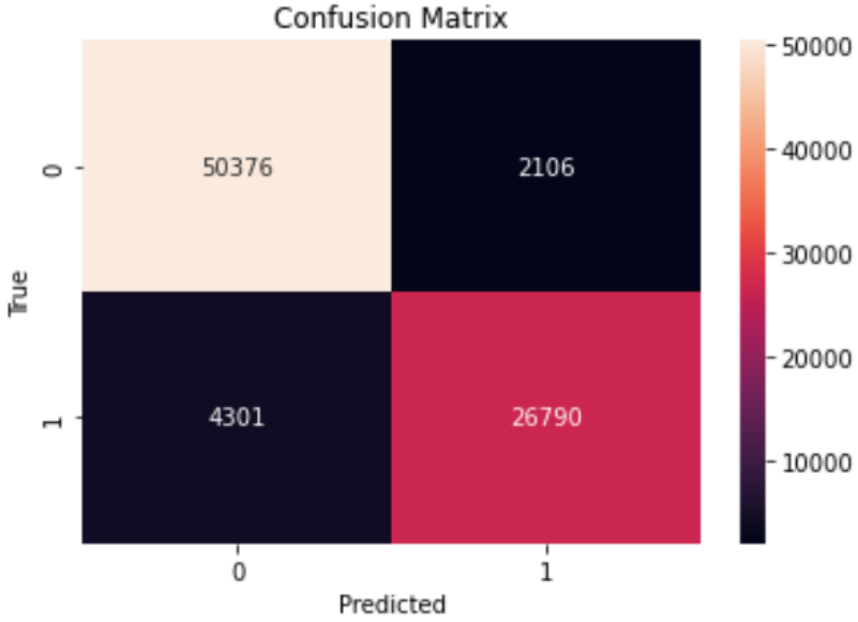
a. True Negatives (TN): 50,376 instances where the model correctly predicted the negative class.

b. False Positives (FP): 2,106 instances where the model incorrectly predicted the positive class.

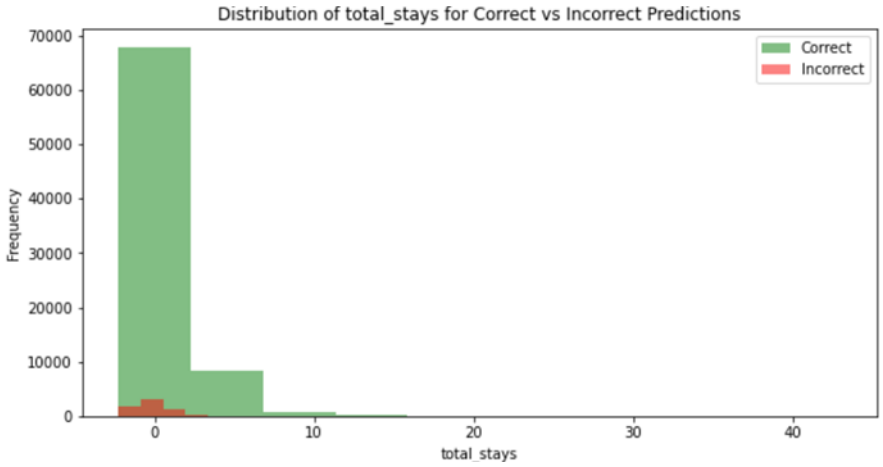
c. False Negatives (FN): 4,301 instances where the model incorrectly predicted the negative class.

d. True Positives (TP): 26,790 instances where the model correctly predicted the positive class.

From this matrix, we can calculate key performance metrics such as accuracy, precision, recall, and the F1 score. Moreover, the relatively high number of false negatives compared to false positives indicates that our model may be more conservative, preferring to predict the negative class. This could be an area of focus for data improvement, possibly by rebalancing the dataset, redefining threshold levels, or further feature engineering to better capture characteristics of the positive class.



2. The majority of correct predictions are clustered around shorter stays. Incorrect predictions also primarily occur with shorter stays but to a lesser extent. This pattern suggests that my model is quite adept at predicting stays of shorter duration. However, the presence of incorrect predictions within the same range indicates that there might be additional factors at play that affect the accuracy of predictions for short stays.



Performance Comparison

I calculate metrics such as AUC, accuracy, precision, recall, and F1-score on the validation dataset for both the Week 9 model and the updated model. Compare these metrics to see which model performs better on the validation dataset. The model with the higher AUC on the validation set is considered to have performed better. If the updated model has a higher AUC, it's likely because the data improvements allowed the model to capture the underlying patterns in the data more effectively, reducing overfitting and increasing its generalization capabilities.

Selecting the Final Model for Deployment

I chose the model with the better performance metrics on the validation dataset as my final model for deployment. This choice is based on the model's ability to generalize to unseen data while minimizing the chance of future performance degradation. Evaluate the final model on the test dataset to get performance metrics. These metrics are critical because they give me the final verification of how my model is expected to perform in the real world.

Ideally, the test error should be comparable to the validation error, which would indicate that the model is generalizing well. A much higher test error could suggest that the model is not as robust as the validation error indicated.

```
Training Accuracy: 0.9233364842712359
Validation Accuracy: 0.8701139155684611
Test Accuracy: 0.8705120330560053
```

Insights Based on Errors:

Potential Overfitting: Given that most of the hotel bookings are likely for shorter durations, the model may be overfitting to the most common scenario in the training data. This could lead to less reliable predictions for less common scenarios, such as longer stays, indicating a need for a more balanced or diversified training set.

Feature Relevance: The errors predominantly occur in the range where most data points lie, which could point to a need for additional features that can help differentiate between classes in this dense area. The creation of new features or the transformation of existing ones might help to reduce these errors.

Data Quality and Engineering: The distribution of errors prompts a review of the data quality, particularly for the cases of longer stays. It may be beneficial to investigate the specifics of these bookings and ensure that all relevant information is being captured. For example, long stays could be associated with different booking behaviors or customer types, which the model may currently overlook.